The Possibilities as a Prediction Tool for Cancer Research of Big Data: Comparison of Incidence Rate of Korean Major Male and Urologic Cancers and Trend Scores

Dong Hyuk Kang[1], Kang Su Cho[2], Won Sik Ham[3], Young Deuk Choi[3], Joo Yong Lee[3]

[1]Department of Urology, Yangpyeong Health Center, Yangpyeong, [2]Department of Urology, Gangnam Severance Hospital, Urological Science Institute, Yonsei University College of Medicine, [3]Department of Urology, Severance Hospital, Urological Science Institute, Yonsei University College of Medicine, Seoul, Korea

Purpose: To examine the trend, and investigated the possibilities as a prediction tool by choosing the trend score about male and urologic cancers have the high incidence rate.

Materials and Methods: We selected 5 major male and 3 major urologic cancers for past 3 years (from 2010 to 2012) and examined the incidence rate, and using Naver and Google trend, the rate of cancer incidence was compared with the trend score during the same period.

Results: From 2010 to 2012, the greatest occurrence of the cancer to males was the stomach cancer, followed by colon, lung, liver, and prostate cancer. In the urologic field, the prostate cancer was the first one, followed by kidney and bladder cancer. In 2010 to 2012, the Naver trend score was 32 for stomach and colon cancers, 31 for lung cancers, 20 for liver cancers, and 19 for prostate cancers, which index were corresponded with the order of incidence rate. Though the Google trend score for prostate cancer was not found, the average was 9 for stomach cancer, 8 for colon cancer, 6 for lung cancer, 4 for liver cancer, which index were corresponded with the order of incidence rate. In 2013 and 2014, the figure of prostate cancer was grown and exceeded liver cancer.

Conclusions: In the trend score, the index of the prostate cancer shows continuing increase, and, from the results, urologists should recognize the importance of the study on the prostate cancer such as management, prevention, and treatment of the prostate cancer. **(Korean J Urol Oncol 2015;13:35-42)**

Key Words: Neoplasms, Urology, Male

INTRODUCTION

Recently, with the expansion in internet, personal computers (PCs), and mobile industries, the digital economy has grown rapidly, so the number of data has increased exponentially. The scale of these data is huge, and the data has a short production cycle and comes in many forms, hence we call the data as the 'Big Data'.[1] According to Doug Laney, an industry analyst, the big data is defined as three factors: 'volume, velocity, and variety'. Aside from these three factors, currently, it is also added 'variability and complexity' to describe as the big data.[2]

Lately, with the social issues of utilizing the big data, it has gained increasing importance. This is because its effective applications are able to produce novel knowledge and to create social and economic values. Actually, the big data is being used

Received March 30, 2015, Revised April 10, 2015 (1st), April 15, 2015 (2nd), Accepted April 16, 2015

Corresponding Author: Joo Yong Lee, Department of Urology, Severance Hospital, Urological Science Institute, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 120-752, Korea. Tel: 82-2-2228-2320, Fax: 82-2-312-2538, E-mail: joouro@yuhs.ac

lot in our everyday lives, for instance, there are representative successful cases such as campaign strategies and predictions on the election in U.S. and Korea,[3,4] advertising systems of companies, raising in investment returns, and crime predictions, and so on. In the health and medical fields, the cases that use the big data are continuing to grow, especially, providing services of flu trends (http://www.google.org/flutrends), Google (Google Inc., Mountain View, CA, USA) has been a considerable achievement in the earlier prediction of diffusion path by figuring out the entering frequency of searched keywords with regard to the flu such as 'symptoms of the flu' and 'treatment of the flu' from region to region all over the world.[5] McKinsey described that the big data has a great expected effect due to a good accessibility of the data in the health and medical fields and a great deal of weight on the economy.[6]

Trend research provides the frequency of searched keywords for certain periods and demographic information about users of search-engines such as Naver (NAVER Corp., Seongnam, Korea), the largest portal site in Korea, and Google, a worldwide search engine, which offers trend research services to public users.[7,8] The search frequency is an index to show the popularity and interest of about user groups' keywords. Although the trend research method is possible to perform simply, the method has its own advantages to analyze and predict the customer demand or social phenomenon accurately.

In today's society, though the incidence rate of cancer has been increased, which rouses the public interest, there is lack of application of big data to the cancers. Besides, the incidence rate of cancer is changing according to the times. Although Korea has a shorter calculation period than other countries that compile the cancer statistics, statistical report may take about 2 years due to the comparison process to check every data against the reported cancer data for avoiding an omission or duplication when Statistics Korea has the report a case that cause of death is a cancer or when Health Insurance Review and Assessment Service (HIRA) has a request medical expense for the cancer. Therefore, in Naver trend and Google trend, the authors chose the trend score about male and urologic cancers which have the high incidence rate, examined the trend, and investigated the possibilities as a prediction tool.

MATERIALS AND METHODS

1. Study design

Based upon the National Cancer Registration and Statistics in 2012, the authors selected 5 male cancers and 3 urologic cancers which have the highest number of occurrences for past 3 years (from 2010 to 2012) and examined the incidence rate, and using Naver trend and Google trend, the rate of cancer incidence was compared with the trend score by checking the trend score during the same period. The incidence rate was used the standardized incidence rate. The date to carry out the search was February 1st 2015, and the search period was set three years (2010-2012) from that based on the ending 2012 which was the last time to conduct the counting by Cancer Registration and Statistics. In addition, the trend score was analyzed by reset the search period on a yearly basis, in 2013 and 2014.

2. Data assessment

1) Cancer incidence rate: During 3 years from 2010 to 2012, using National Cancer Registration and Statistics (http://ncc.re.kr/manage/manage03_033_list.jsp) in 2012, the authors obtained the overall average of both males and females for the standardized incidence rate of 5 major cancers that had more occurred in males and calculated the overall average of both males and female for the standardized incidence rate of 3 major urologic cancers that had more happened in urologic areas.

2) Naver trend: After accessing the Naver trend (http://trend.naver.com/), the authors searched the trend of the 5 major male cancers and 3 major urologic cancers. The classification was chosen the 'PC' between 'PC' and 'mobile', three periods were selected, from January 2010 to December 2012, from January 2013 to December 2013, and from January 2014 to December 2014. Because the trend score is not the absolute index of searched amounts, but the relative index depending on the flows, the author also calculated the calibrated index were expressed as a percentage terms, of which the highest trend score of each period is 100, for retaining objectivity.

3) Google trend: After accessing the Google trend, the authors searched the trend of the 5 major male cancers and 3 major urologic cancers. In consideration of searching with Korean language, the national division was selected as 'worldwide'.

The category was chosen the 'All category', the classification was selected as 'web search' among 'web search', 'Image search', 'News search', 'Google shopping, and 'YouTube search'. Three periods were selected, from January 2010 to December 2012, from January 2013 to December 2013, and from January 2014 to December 2014. Because the trend score is not the absolute index of searched amounts, but the relative index depending on the flows, the authors also calculated the calibrated index were expressed as a percentage terms, of which the highest trend score of each period is 100, for retaining objectivity.

RESULTS

1. 5 major male cancers and 3 major urologic cancers in Korea

For three years from 2010 to 2012, the greatest occurrence of the cancer to males in Korea was the stomach cancer (62,362 people), followed by colon cancer (50,214), lung cancer (45,184), liver cancer (36,159), and prostate cancer (26,342) came in from second to fifth as 5 major cancers. In the urologic field, the prostate cancer (26,342) was the first one, followed by kidney cancer (11,845) and bladder cancer (10,557) came in second and third according to the number of occurrences.

2. Comparison of the overall average of both males and females for the standardized incidence rate and trend scores of 5 major male cancers

1) In 2010-2012: For three years from 2010 to 2012, the overall average of both males and females for the standardized incidence rate of 5 major male cancers was 44.2 for stomach cancers, 38.0 for colon cancers, 29.1 for lung cancers, 23.4 for liver cancers, and 11.5 for prostate cancers. During the same period, the overall average of trend score (calibrated index) extracted by Naver trend was 32 (100) for stomach cancers, 32 (100) for colon cancers, 31 (96.9) for lung cancers, 20 (62.5) for liver cancers, and 19 (59.4) for prostate cancers, which index were corresponded with the order of standardized incidence rate (Fig. 1A). Though the overall average of the trend score extracted by Google trend for prostate cancer was not found, the overall average was 9 (100) for stomach cancer, 8 (88.9) for colon cancer, 6 (66.7) for lung cancer, 4 (44.4) for liver cancer, which index were corresponded with the order of standardized incidence rate (Fig. 2A) (Table 1).

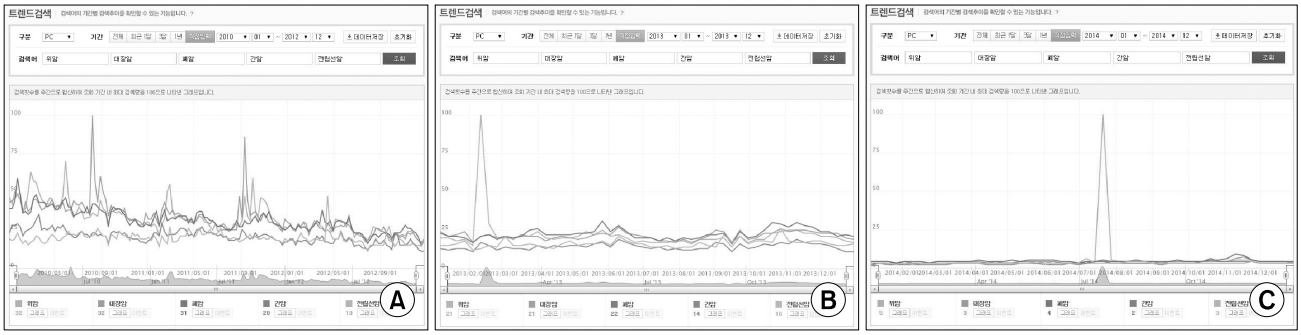

Fig. 1. Interest over time on 5 major male cancer in Korea using Naver Trend (A) in 2010-2012, (B) in 2013, (C) 2014.

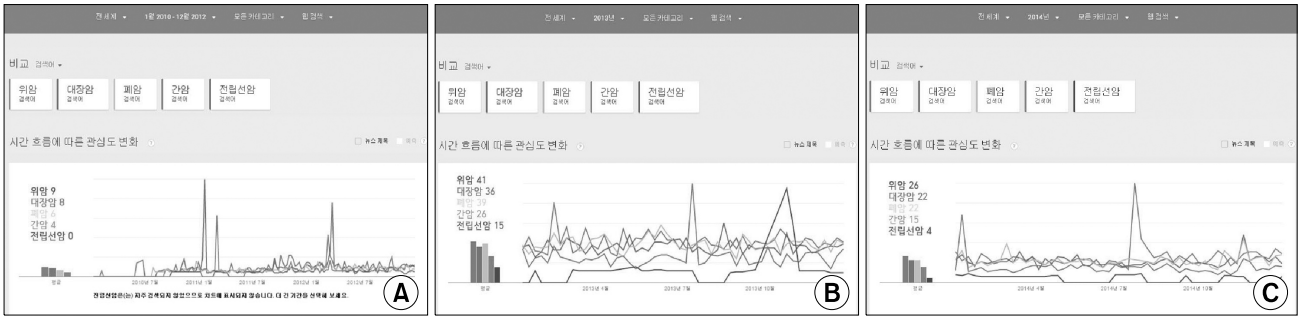

Fig. 2. Interest over time on 5 major male cancer in Korea using Google Trend (A) in 2010-2012, (B) in 2013, (C) 2014.

Table 1. Standardized incidence rate and interest using Naver and Google Trend over time on 5 major male cancer in Korea

	2010~2012										2013						2014					
	Standardized incidence rate				Naver trend			Google trend			Naver trend			Google trend			Naver trend			Google trend		
	2010	2011	2012	Mean	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank
Gastric cancer	44.2	44.4	41.2	43.3	1	32	100	1	9	100	1	21	95.5	2	41	100	1	5	100	1	26	100	1
Colon cancer	38.0	39.4	38.6	38.7	2	32	100	1	8	88.9	2	21	95.5	2	36	87.8	3	3	60	3	22	84.6	2
Lung cancer	29.1	29	27.9	28.7	3	31	93.9	3	6	66.7	3	22	100	1	39	95.1	2	4	80	2	22	84.6	2
Liver cancer	23.4	23.0	21.7	22.7	4	20	62.5	4	4	44.4	4	14	66.7	5	26	63.4	4	2	40	5	15	57.7	4
Prostate cancer	10.9	11.9	11.6	34.4	5	19	59.4	5	-	-	5	16	76.2	4	15	36.6	5	3	60	3	4	15.4	5

*Calibrated index were expressed as a percentage terms, of which the highest trend score of each period is 100, for retaining objectivity.

2) in 2013: The lung cancer ranked top in Naver trend score, and the figure of prostate cancer was grown and exceeded the score of liver cancer (prostate cancer 16 (76.2) vs liver cancer 14 (66.7)) (Fig. 1B). As increasing the figure of the lung cancer, it came in ahead of the colon cancer to gain second place in Google trend score, and the prostate cancer that has no index due to the small quantity to search scored 36.6 of calibrated index, but it came fifth (Fig. 2B) (Table 1).

3) In 2014: In Naver trend score, the stomach cancer ranked again top, and the lung cancer scored second place. Following the trend score of prostate cancer in 2013, its score was higher than the score of liver cancer, and the same score as the colon cancer. (prostate cancer 3 (60) vs liver cancer 2 (40) vs colon cancer 3 (60)) (Fig. 1C). In Google trend score, the stomach cancer ranked top, the colon cancer and lung cancer were recorded in same figure (Fig. 2C) (Table 1).

3. Comparison of the overall average of both males and females for the standardized incidence rate and trend scores of 3 major urologic cancers

1) In 2010-2012: For three years from 2010 to 2012, the overall average of both males and females for the standardized incidence rate of 3 major urologic cancers was 11.5 for prostate cancers, 5.83 for kidney cancers, 4.63 for bladder cancers. During the same period, the average trend score extracted by Naver trend (calibrated index) was 41 (100) for prostate cancers, 18 (43.9) for kidney cancer, and 20 (48.8), which did not have a great difference though the figure of bladder cancer was higher than the kidney cancer compared to the standardized incidence rate (Fig. 3A). Similarly, the average trend score (calibrated index) extracted by Google trend was 43 (100) for prostate cancer, 16 (37.2) for kidney cancers, and 18 (41.9) for bladder cancers, which did not have a great difference though the figure of bladder cancer was higher than the kidney cancer compared to the standardized incidence rate (Fig. 4A) (Table 2).

2) In 2013: In Naver trend score, the prostate cancer ranked top, and the kidney cancer and bladder cancer had a same index (kidney cancer 25 (40.3) vs bladder cancer 25 (40.3)) (Fig. 3B). On the other hand, in Google trend score, the prostate cancer was only searched, and the kidney cancer and bladder cancer did not be found (Fig. 4B) (Table 2).

3) In 2014: In Naver trend score, the index was 50 (100) for prostate cancers, 16 (32) for kidney cancers, 19 (38) for bladder cancer, which did not have a great difference though

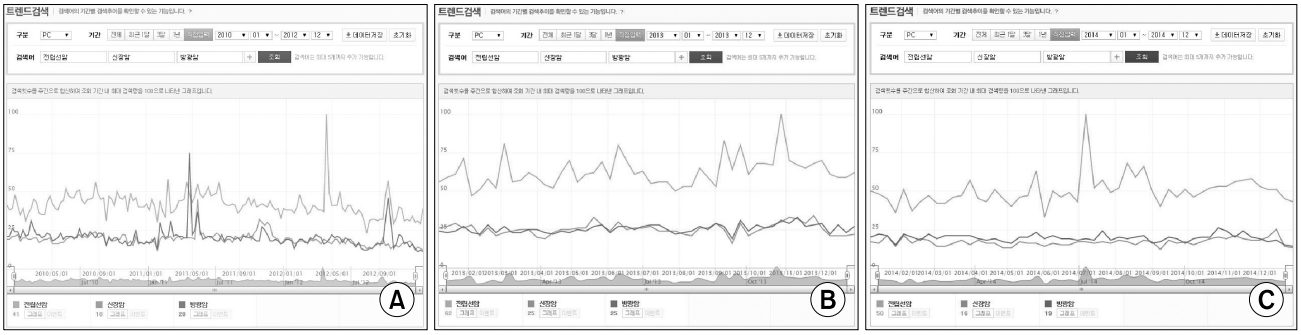

Fig. 3. Interest over time on 3 major urologic cancer in Korea using Naver Trend (A) in 2010-2012, (B) in 2013, (C) 2014.

Fig. 4. Interest over time on 3 major urologic cancer in Korea using Google Trend (A) in 2010-2012, (B) in 2013, (C) 2014.

index of bladder cancer was higher than the kidney cancer (Fig. 3C). On the other hand, in Google trend score, only the trend of the prostate cancer was searched, and the kidney cancer and bladder cancer did not be found (Fig. 4C) (Table 2).

DISCUSSION

For the last several decades, the top cause of death has been the cancer. According to the cause of death statistics, total death toll is 266,257 people in 2013, of these, 75,334 people that is about 28.3% died of cancers.[9] Cancer Registration and Statistics Program is a project to widely utilize in policy development, direction, evaluation, cancer research, and so on of the national cancer management project, and it will provide reliable and accurate statistical data of the cancer that is top cause of death and will precisely and timely calculate the national cancer statistics.[10] In 1980, 'Korean Multi-Hospital Central Center Registry' at the National Medical Center in Korea was installed and operated, then it have reported results of the survey project for registered cancers on a yearly basis. This project transferred to the National Cancer Center, and it was designated as Multi-Hospital Central Center Registry by Ministry of Health & Welfare in 2004, thus, Cancer Registration and Statistics Program is under way.

Domestic interests in the cancer, cancer management projects, and cancer treatments have taken on added importance. Recently, the big data is tended to make the possible use in the fields of cancer researches and treatments. Though the projects are still in its infancy, the possibilities of the big data are unlimited. Using Naver trend and Google trend which are relatively easy to access as representative big data, the authors have pursued the study to compare with the incidence rate of cancers and to investigate the application possibilities of the big data to the male cancers and urologic cancers. As a result, the standardized incidence rate of locally major male cancers was corresponded with the ranking of trend scores from Naver and Google, and it showed a higher match rate relatively when comparing with the calibrated index. Searching trend is not to confirm absolute searching quantities from the number of searches, but to identify the searching patterns, which reflects the change in interests of web users. With a high conformity degree between the standardized incidence rate and the trend score, the authors could confirm the using possibility as the big data to be able to predict the project of cancer statistics.

Table 2. Standardized incidence rate and interest using Naver and Google Trend over time on 3 major urologic cancer in Korea

	2010~2012										2013						2014					
	Standardized incidence rate				Naver trend			Google trend			Naver trend			Google trend			Naver trend			Google trend		
	2010	2011	2012	Mean	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank	Score	Calibrated index*	Rank
Prostate cancer	10.9	11.9	11.6	11.5	1	41	100	1	43	100	1	62	100	1	100	100	1	50	100	1	100	100	1
Kidney cancer	5.6	6.0	5.9	5.8	2	18	43.9	3	16	37.2	3	25	40.3	2	-	-	-	16	32	3	-	-	-
Bladder cancer	4.8	4.7	4.4	4.6	3	20	48.8	2	18	41.9	2	25	40.3	2	-	-	-	19	38	2	-	-	-

*Calibrated index were expressed as a percentage terms, of which the highest trend score of each period is 100, for retaining objectivity.

Specifically, in 2013 and 2014, it could be confirmed that Naver trend score reverted in liver cancers that tend to be continued to decrease the incidence rate (27.3 → 21.7, 2002 → 2012) and prostate cancers that tend to be continued to increase the rate (3.9 → 11.6, 2002 → 2012). It is too early to conclude that the incidence rate of the prostate cancer would be able to predict the indirect increase, and urologists should further recognize the importance of research on the prostate cancer such as management and prevention projects and treatment of the prostate cancer based on the prediction. In particular, this study has significant meaning in terms of an attempt to predict the relation to incidence rate of diseases with the big data that is readily accessible.

The rapid development of IT technologies has increased the absolute quantities of the data to be analyzed and processed. On July, 2008, as launching iPhone 3G (Apple Inc., Cupertino, CA, USA) that is a representative smart phone, the genuine mobile revolution was started, and as interworking with a social network service (SNS), its users increased dramatically, thus, the production rates of generated data has further gained speed.[11] Gartner (Stamford, CT, USA), a market survey institutes of technology information, reported that the big data was examined as top 10 of future strategy and technology in 2012.[12] McKinsey suggested the five fields to be able to be utilized the big data; medical and health, public administration, personal information like a location, retail and distribution, and manufacturing fields,[6] practically, all over that world, governments and corporations are using the big data or preparing to use them.

There have been many cases to utilize the big data in the medical field. WellPoint (Anthem Inc., Indianapolis, IN, USA) that is a health insurance company is assisting to search complex medical treatments based on the integrated analysis of 3,420 patients' information who are registered in the company. This big data contains medical insurance claims data, patients' demographic information, history of medical service use, history of prescription, image data, and clinicopathological data as well as all information about treatment history such as patients' symptom, interview result, and diagnosis.[13] In case of The National Health Service (NHS) in UK, the big data has been utilized various ways to carry out forecasting the national health to make a database (DB) with prescription data of pharmacies and hospitals across the nation. In addition, with a website, Clinical Practice Datalink (CPRD), interworking various dataset

is provided to researchers.[14] Also, to predict diseases, there are diverse using cases of the big data. The flu trend provided by Google (http://www.google.org/flutrends/) is a representative example to use the keyword data. Using the searched keywords of Google, it shows the patterns by processing certain searched keywords as a service to predict the sequence of flu endemic in real-time. Though the all people who search the word of flu are not sick, it was confirmed that related keywords tends to increase in flu endemic, so the authors could have foreseen how endemic the flu is by calculation of the frequency of the searched keywords. Also, visualizing the endemic degree of the flu in every region with a location based system, the information of the region that holds a high risk of flu is provided.[5] Although there is much to be desired as a system, compare to other health organizations that updates the data a week, Google updates the information of the flu endemic on a daily basis, which is regarded as enough role in complement to existing systems. In USA, using the twitter, a SNS, forecasting system of diseases was developed, which realized the possible technology to track the diverse diseases from influenza to allergy.[15]

Likewise, though the utilization of big data still needed the expansion depending on the overall trends in medical view, its applications in the fields of cancer researches and treatments is looming. Angelina Jolie, a famous movie star, became an issue due to the operation of the preventive double mastectomy, predicting the rates of breast cancer through her own gene test with a family history.[16] The operation showed you how huge people's belief on the future predicted by the big data really is. IBM (International Business Machines Corp., Armonk. NY, USA) and Memorial Sloan-Kettering Cancer Center (MSKCC) located in New York entered into a business agreement to help doctors to choose the diagnostic and treatment methods for cancers through data mining which finds novel knowledge based on the mass data of Watson, a super computer.[17] In Korea, the movement to use the big data for cancer research is also beginning. For example, National Cancer Center is starting to carry out the construction project of the big data DB from this year and trying to conduct the numerous reaches on the bases of these DB. These researches has a meaningful achievement to examine the possibilities as a tool to predict the cancer incidence rate in advance by grasping relations with incidence rate with an easily accessible data. Of course, it is obvious that is has many faults and risks to provide erroneous predictions. However, we think that these big data are available to use as supplemental tool for prediction in the field of cancer statistics that takes two years through a detailed and systematic study.

There were several limitations for the application of big data to the medical field, including this research. First, the searching the trend, which was conducted by this study, is affected greatly by the search frequency in a certain period because it does not provide absolute quantities of searching, but provides relative quantities depending on the passing of time. Thus, because the hitting number is abnormally increased in case of unique events such as death of celebrity or development of medicine, the fact that it is difficult to judge objectively leaves much to be desired. Thus it is very important issue how to interpret the result and we should judge cautiously the value of study. There is a high probability to cause problems not to be foreseen because nothing has been actively introduced. Currently, because of a lot of untested parts, having blind faith in the big data might act as venom. Because ignoring existing epidemiologic methods by falling in the data analysis might be sticking your neck out, we think that seeking ways to use as a supplemental method to the existing systems is one answer. Besides, the information security still remains as an unsolved problem because the big data is commonly easy to leak the private information.

It, however, is clear that there is an unlimited potential to be developed. If greater use were made of the big data, the cause of diseases that has never known before can be found, or the rates of diseases that are difficult to predict can be forecasted, furthermore, we think that it will be great help to cure the diseases and improve the quality of patients' life. In the cancer field, particularly, the big data is an insubstantial field and has a high probability of development, which is necessary to further analyze and research. Although there is coexistence of both expectations and concerns, all doctors, patients, and relative professions of medical devices or IT should positively cooperate and work out even better results by minimization of problems and maximization of the positive aspects.

CONCLUSIONS

The standardized incidence rate of major male cancers in Korea score is consistent with the ranking order of Naver and Google trend. In the trend score, the index of the prostate cancer shows continuing increase, and, from the results, urologists should recognize the importance of the study on the prostate

cancer such as management, prevention project, and treatment of the prostate cancer. Though there is much to be desired, we think that these big data are available to be used as a supplemental tool in cancer research projects through detailed and systematic studies.

REFERENCES

1. McAfee A, Brynjolfsson E. Big data: the management revolution. Harvard Business Review, 2012;90:60-6,68,128
2. Laney D. 3D Data management: controlling data volume, velocity and variety. Gartner. Retrieved 2011;6
3. Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. Proceedings of the ACL 2012 System Demonstrations:115-20
4. Bae J, Son J, Song M. Analysis of twitter for 2012 South Korea presidential election by text mining techniques. Journal of Intelligence and Information Systems 2013;19:141-56
5. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One 2011;6: e23610
6. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute 2012
7. Naver trend, accessed http://trend.naver.com
8. Google trend, accessed https://www.google.co.kr/trends/?hl=ko
9. The causes of death statistics of Korea 2013, Statistics Korea, accessed http://kostat.go.kr/portal/korea/kor_nw/2/1/index.board?bmode=read&aSeq=330181
10. Cancer Registration and Statistics Program, National Cancer Canter, accessed http://ncc.re.kr/manage/manage12_01.jsp
11. Lee JY, Kang DH, Moon HS, Kim YT, Yoo TK, Choi HY, et al. Analysis of content legibility for smartphones of websites of the Korean urological association and other urological societies in Korea. Korean J Urol 2011;52:142-6
12. Prentice S. CEO advisory: “Big Data” equals big opportunity. Gatner. 2011;Research ID: G00211628
13. Groves P, Kayyali B, Knott D, Van Kuiken S. The ‘big data’ revolution in healthcare. McKinsey Quarterly, 2013
14. Kim J, Kim H, Sohn G, Song Y, Yoon J, Lim H, et al. Big data and medicine. Communications of KIISE 2014;32:18-26
15. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2011;8:e83672
16. Jolie A. My medical choice. The New York Times, 2013 accessed http://www.nytimes.com/2013/05/14/opinion/my-medical-choice.html?_r=0
17. Gales J. Watson turns medic: Supercomputer to diagnose disease. New Scientist 2012;215:19